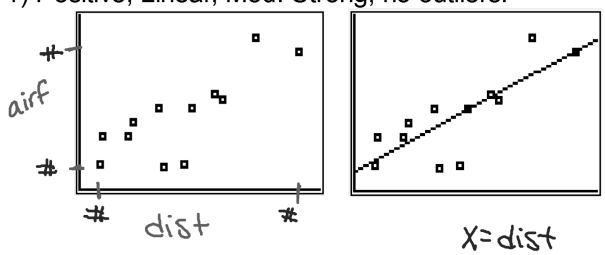
Answers to worksheet:

1) Positive, Linear, Mod. Strong, no outliers.



2) $\widehat{airfare}$ = 83.267 + 0.1174(distance)

r = 0.795 r^2 = 63.2%

3) airfare = 83.267 + 0.1174(370) Y1(370)

airfare = $126.70

residual = $138 -- $126.70 = $11.30

4) underestimate (prediction < actual)

5) airfare = 83.267 + 0.1174(2842)

airfare = $416.85

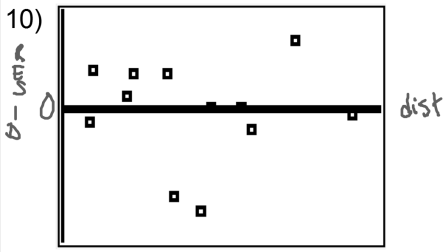
6) NO! Extrapolation. 2842 would be an outlier in the X-variable.

7) For every 1 mile traveled on a flight, the airfare increases by $0.1174 (or $0.12).

8) $11.74 or $12.00

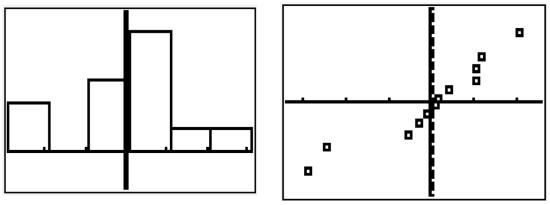
9) 63.2% of the change in airfares is explained by the change in distance flown.

10)



11) Residual plot = scattered, therefore the line is a good model for our data. (Also, the r and r^2 are high and the original scatterplot was linear.)

12)



residuals normally distr. RESID

More review:

- Given the following regression output, what is the LSRL equation? the correlation?

Dependent variable is: No Opinion (Y)
R-squared = 9.5%
s = 2.280 with 16 − 2 = 14 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept b_0	7.69262	2.445	3.15	0.0071
X Year b_1	−0.042708	0.0353	−1.21	0.2458

no opinion = 7.69262 − 0.042708(year)

r^2 = 0.095 r = −0.3082

LinReg T-Test

* We want to test if there is an association between two **quantitative** variables (x and y)

* We look at a SAMPLE of data on a scatterplot and estimate the true relationship of the POPULATION plot

* LSR line (sample): $\hat{y} = b_0 + b_1 x$

Where: $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$

* Population model (that we don't know): $y = \beta_0 + \beta_1 x + \varepsilon$ $\beta = 0$ $\beta > 0$

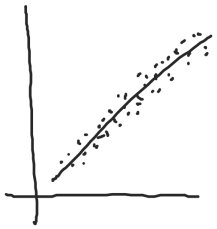
SAMPLE	POPULATION	WHAT IS IT?
b_0	β_0	intercept
b_1	β_1	slope
e_i	ε_i	errors (residuals)

Residuals (errors):

* Independent, Normally distributed
* N(0, S_e)

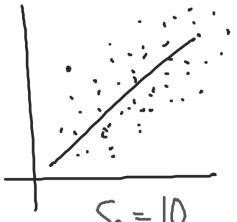
$$s_e = \sqrt{\frac{\sum e^2}{n-2}}$$

small S_e of residuals:



$S_e = 2$

large S_e of residuals:



$S_e = 10$

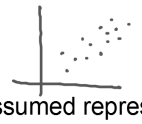
$Resid = N(0, S_e)$

T-Test: Testing the slope of the population regression line

Hypotheses:

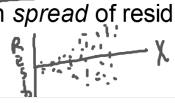
Ho: $\beta_1 = 0$ (there is no association btw X & Y)

Ho: $\beta_1 >, <, \neq 0$ (there is an association btw X & Y)



Conditions:

1) SRS	1) stated SRS or assumed represent.
2) Linear data	2) scatterplot is linear w/ no outliers
3) Independence	3) each piece of data can be assumed indep. of the others
4) Normal residuals	4) Normal prob. plot of residuals is linear LRESID
5) Equal variance	5) Residual plot shows no change in *spread* of residuals

NO: 

Conditions met --> t-distribution --> LinReg t-test

Ho: $\beta_1 = 0$

Ha: $\beta_1 = 1$

Mechanics:

Test Statistic:

$$t = \frac{\text{statistic - parameter}}{\text{std. dev. of stat (SE)}} = \frac{b_1 - \beta_1}{SE_{b_1}}$$

If $\neq$, 2x

P-Value: P(t $\gtrless$ test statistic) = tcdf(LB, UB, df)

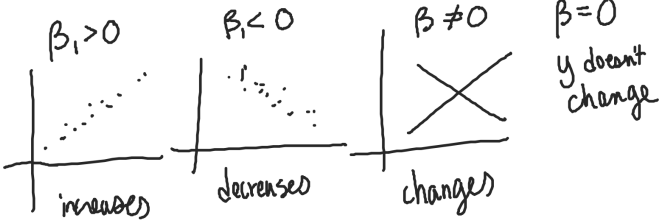
df = n - 2

Conclusion: (3 sentences)

* Reject / Fail to reject Ho ....

* We have sufficient/insufficient evidence that the slope of the population regression line is _____ 0.

* Thus as _X_ increases, _Y_ _____.



2 ways to do the mechanics of this test:

1) With actual data
   * Use LinReg t-test

Example: Airfare Data

L_1 L_2

$S \neq SE_b$

S = std. dev. of errors.

$H_0: \beta_1 = 0$

$H_a: \beta > 0$

cond✓

df = 10

$t = \frac{0.1174}{0.0283} = 4.144$

$P(t > 4.144) = 9.993 \times 10^{-4}$

2) With computer output:

Model of HusbandsAndWives — Multiple Regression

Response attribute (numeric): Age_Wife

$n = 40$

Predictor	Coefficient	Std Error	t Statistic	P Value	ΔR^2
Constant	b_0 1.5740	[illegible]			
Age_Husband	b_1 0.9112	0.0259 (SE_{b_1})	35.249	0.0000	0.8809

Regression Equation: $\widehat{Age\_Wife}$ = 1.57400798388 + 0.911241590288 Age_Husband

R-Squared: **0.880894**
Adjusted R-Squared: **0.880186**
Standard Deviation of the Error: **3.95101**

$$t = \frac{b_1 - \beta_1}{SE_{b_1}} = \frac{0.9112}{0.0259}$$

$P(t > 35.249)$

Example:

1) Does a relationship exist between High School GPA and freshman year performance in college? A random sample of 40 freshman at a local college was taken and their HS GPA and the GPA from their first full year were recorded.

$n = 40$

Hypotheses:

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Model of Sample of SATGPA — Multiple Regression

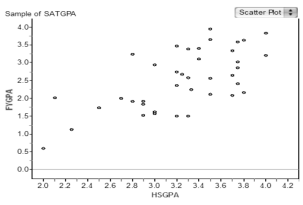
Response attribute (numeric): FY GPA

Predictor	Coefficient	Std Error	t Statistic	P Value	ΔR^2
Constant	b_0 -1.0149	[illegible]			
HSGPA	b_1 1.0903	0.2045 (SE_{b_1})	5.333	0.0000	0.4280

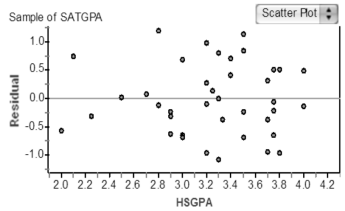
Regression Equation: $\widehat{FY\ GPA}$ = −1.01493878794 + 1.09034493208 HSGPA

R-Squared: **0.428036**
Adjusted R-Squared: **0.412985**
Standard Deviation of the Error: **0.626502**

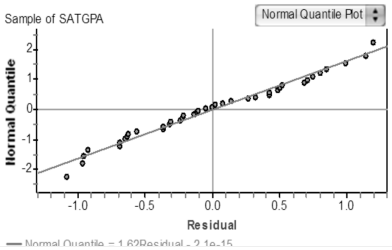
Scatterplot



Residual Plot



Normal Probability Plot



Conditions:

1) SRS	1) stated
2) Linear data	2) orig. plot is ≈ linear
3) Indep.	3) each coll. freshman is indep. of each other
4) Normal resid.	4) the normal prob. plot of resid is linear
5) Equal var.	5) resid. plot is equally spread

Test Statistic:

$$t = \frac{b_1 - \beta_1}{SE_{b_1}} = \frac{1.0903 - 0}{0.2045} = 5.332$$

P-Value:

$2 \cdot P(t > 5.332) = 4.687 \times 10^{-6}$

$df = n - 2 = 38$

Conclusion:

- We reject H_0 b/c p-value of $4.687 \times 10^{-6} < \alpha = 0.05$.
- We have sufficient evidence that the slope of pop. regr. line is $\neq 0$.
- Thus, as HSGPA increases, college GPA changes.

Confidence Interval:

Formula: statistic ± (critical value)(SE) = (a, b)

$b_1 \pm (t^*)(SE_{b_1})$

- Get the t* from the INVT program.

- Be careful!
  * If sample size in the problem is 40, then df is n - 2 = 38
  * So in the INVT prgram you have to put in n = 39.
  * The program assumes that the df is n - 1.

Sentence:

We are ____% confident that as __X__ increases by 1 X unit, the __Y__ increases/decreases between __a__ and __b__ Y units.

Example: AIRFARES data (n = 12) 95% confidence

(0.1174) ± (2.228)(0.0283) = (0.05435, 0.1804)

We are 95% confident that as the distance increases by 1 mile, the airfare increases between $0.05435 and $0.1804.

Book problems: p. 674 #5, 13

5) (a) Conditions:

1) SRS	1) assume 2005 movies are representative of all movies
2) Linear Data	2) The scatterplot looks roughly linear with 1 possible outlier
3) Independence	3) Each movie is indep. of others
4) Normal residuals	4) The normal probability plot of the residuals looks approx. linear --> normal residuals
5) Equal Variance	5) The residual plot shows no change in the spread of the residuals.

conditions met --> t distribution --> Lin Reg T-test

(b) 0.7144 ± (1.98)(0.1541) = (0.4094, 1.0194)

We are 95% confident that for every increase of 1 minute of run time of a movie, the budget increases btw 0.4094 and 1.0194 million dollars.

13) (a) Ho: $\beta_1 = 0$
Ha: $\beta_1 < 0$

(b) No, the conditions are not satisfied because the histogram is not normally distributed. It is right skewed

(c) $t = \frac{-0.02996}{0.0043} = -7.04$

$P(t < -7.04 \mid df = 26) = 8.907 \times 10^{-8}$

We reject Ho b/c pvalue of $8.907 \times 10^{-8} < \alpha = 0.05$.
We have sufficient evidence that the slope of the pop. regr. line between year and difference in marital ages of men & women is less than 0. Therefore as the year increases the difference between marital ages decreases.

HW: p. 674

#4 (not b or e)

6

18

35 (for b, just do a regular 95% conf. int)